dbCerEx: A Web-Based Database for the Analysis of Cervical Cancer Transcriptomes

Limin Zhou[1], Wei Zheng[2], Majing Luo[4], Jing Feng[5], Zhichun Jin[1], Yan Wang[1], Dunlan Zhang[1], Qiongxiu Tang[1], Yan He[3]*

1 Hubei Maternal and Child Health Hospital, Wuhan, Hubei, P.R. China, **2** Yichang Humanwell Pharmaceutical Co.,Ltd, Yichang, Hubei, P.R. China, **3** College of Fisheries, Huazhong Agricultural University, Wuhan, P.R. China, **4** College of Life Sciences, Wuhan University, Wuhan, P.R. China, **5** International school of software, Wuhan University, Wuhan, P.R. China

Abstract

Background: Cervical cancers are ranked the second-most hazardous ailments among women worldwide. In the past two decades, microarray technologies have been applied to study genes involved in malignancy progress. However, in most of the published microarray studies, only a few genes were reported leaving rather a large amount of data unused. Also, RNA-Seq data has become more standard for transcriptome analysis and is widely applied in cancer studies. There is a growing demand for a tool to help the experimental researchers who are keen to explore cervical cancer gene therapy, but lack computer expertise to access and analyze the high throughput gene expression data.

Description: The dbCerEx database is designed to retrieve and process gene expression data from cervical cancer samples. It includes the genome wide expression profiles of cervical cancer samples, as well as a web utility to cluster genes with similar expression patterns. This feature will help researchers conduct further research to uncover novel gene functions.

Conclusion: The dbCerEx database is freely available for non-commercial use at http://128.135.207.10/dbCerEx/, and will be updated and integrated with more features as needed.

Citation: Zhou L, Zheng W, Luo M, Feng J, Jin Z, et al. (2014) dbCerEx: A Web-Based Database for the Analysis of Cervical Cancer Transcriptomes. PLoS ONE 9(6): e99834. doi:10.1371/journal.pone.0099834

Editor: I. King Jordan, Georgia Institute of Technology, United States of America

Received November 12, 2013; **Accepted** May 19, 2014; **Published** June 11, 2014

Funding: This study is supported by the National Natural Science Founds of China (Program No. 31202013), Specialized Research Fund for the Doctoral Program of Higher Education of China (Program No 20120146120007), Fundamental Research Funds for the Central Universities (Program No.2011QC075). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The commercial company (Yichang Humanwell Pharmaceutical Co.,Ltd,), along with any other relating to employment, consultancy, patents, products in development or marketed products etc., have declared that no competing interests and Financial disclosure. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

* E-mail: heyan@mail.hzau.edu.cn

Introduction

Cervical cancers account for the second-most gynecological cancer death cases worldwide, and this situation is worse in developing countries due to the lack of adequate organized screening programs. It is believed that Human Papilloma Virus (HPV) infections are the major causes of invasive cervical cancer [1].

Whole- genome expression profiling has revolutionized in the way we study disease and basic biology. Since 1997, the number of published results based on an analysis of gene expression microarray data has grown from 30 to over 5,000 publications per year [2]. DNA microarray technologies aim at simultaneous measurements of the expression of thousands of genes in one single experiment. Over the past few years, this technology has facilitated better understanding of the complex and heterogeneous molecular characteristics of cancers and helped to improve treatment in cancers. For example, HOXC10 gene at first was identified to belong to the 171 significantly up-regulated genes in the cervical squamous cell carcinomas (SCC) relative to normal cervix samples from DNA microarray, which was later identified as a key mediator of invasion in cervical cancer [3]. Archival RNA samples of 25 patients were hybridized to Stanford microarray chips to build a seven gene scoring system [4]. This gene expression pattern could help to identify patients with cervical cancer who can be treated with radiotherapy alone. The specific expression profiles of candidate genes were selected to identify historical subtypes of cervical cancer [5]. Furthermore, numerous candidate biomarkers and therapeutic targets have been identified in other cancers.

However, for most of the published microarray studies, only subsets of genes have been reported to demonstrate the authors' hypothesis. The complete microarray datasets are stored in an unsystematic manner, and useful only to those with computational expertise. Also, RNA-Seq data has become more standard for transcriptome analysis and is widely applied in cancer studies. While for most of the experimental researchers, there also remain difficulties to utilize these cancer microarray databases and RNA-Seq data to solve biological questions. For example, if one novel gene of interest has a correlated (positive or negative) expression pattern with an apoptosis-related gene, it indicates that they may share the same regulatory mechanism, which could provide the potential research proposal for the novel gene.

Table 1. List of GEO accession number, published year and expression platforms of microarray experiments and RNA-Seq data used in this study.

	GEO Acc.*	Year	Expression Platform	Sample Information	Reference
1	GSE5787	2006	Affymetrix Human Genome U133 Plus 2.0 Array	Sixty-six flash-frozen punch biopsies were obtained from 16 patients with cervical cancer.	[13]
2	GSE3578	2007	GE Healthcare/Amersham Biosciences CodeLink Human Whole Genome Bioarray	Twenty-eight squamous cell carcinoma of cervix from 24 patients were taken as biopsy sample before treatment and during treatment	[14]
3	GSE6791	2007	Affymetrix Human Genome U133 Plus 2.0 Array	84 cervical cancers, head and neck cancers and site-matched normal epithelial samples from 20 patients	[15]
4	GSE10372	2008	Sentrix Human-6 Expression BeadChip	32 snap-frozen tissues from 68 cervical carcinomas patients who underwent radical hysterectomy with bilateral lymphadenectomy between 1991 and 2005.	[16]
5	GSE9750	2008	Affymetrix Human Genome U133A Array	A total of 66 samples were included, which include 33 primary tumors, 9 cell lines, and 24 normal cervical epithelium.	[17]
6	GSE20167	2010	GE Healthcare/Amersham Bioscience CodeLink Human Whole Genome Bioarray	A total of 80 cevical cancer samples of following histology were included in this study: 54 squamous cell carcinoma, 18 adenosquamous carcinomas, 6 adenocarcinoma, and 2 others	[5]
7	GSE29570	2012	Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]	The polymorphism of mtDNA D-Loop was investigated in 187 cervical cancer patients and 270 healthy controls.	
8	GSE39001	2013	Affymetrix Human HG-Focus Target Array	43 HPV16-positive cevical cancer and 12 healthy cervical epitheliums using the HG-Focus microarray	
9	GSE27469	2013	Illumina HumanWG-6 v3.0 Expression beadchip	82 patients with cervical cancer, stage 1b bulky through 4a, were included	[18]
10	TCGA-CESC	2014	RNASeq TCGA	The total number of Cervical squamous cell carcinoma and endocervical adenocarcinoma samples is 190.	[10]

*NCBI Gene Expression Omnibus Accession number, it can be used to retrieve the microarray experiment data via http://www.ncbi.nlm.nih.gov/geo/.
doi:10.1371/journal.pone.0099834.t001

Here we present dbCerEx, a database of gene expression profiles generated from DNA microarray experiments and RNA-Seq data. The database is provided with an integrated web-based utility, which has made the data easily accessible to the cervical cancer research community. According to this method, the experimental researchers could identify novel cervical cancer related genes and explore the relationships among them.

Construction and Content

Microarray and RNA-Seq Data

The microarray expression data (GSE matrix files) and platform annotation (GPL files) were retrieved from Gene Expression Omnibus (GEO) database [6] via a R [7]/Bioconductor [8] 'GEOquery' package [9]. The RNA-Seq data were retrieved from The Cancer Genome Atlas (TCGA) Data Portal [10], which

Table 2. Predefined Gene Sets.

Category	Gene set title	Number of gene sets
Pathway	BIOCARTA	217
	KEGG	186
	REACTOME	674
Gene Ontology	Biological process	825
	Cellular component	233
	Molecular function	396

doi:10.1371/journal.pone.0099834.t002

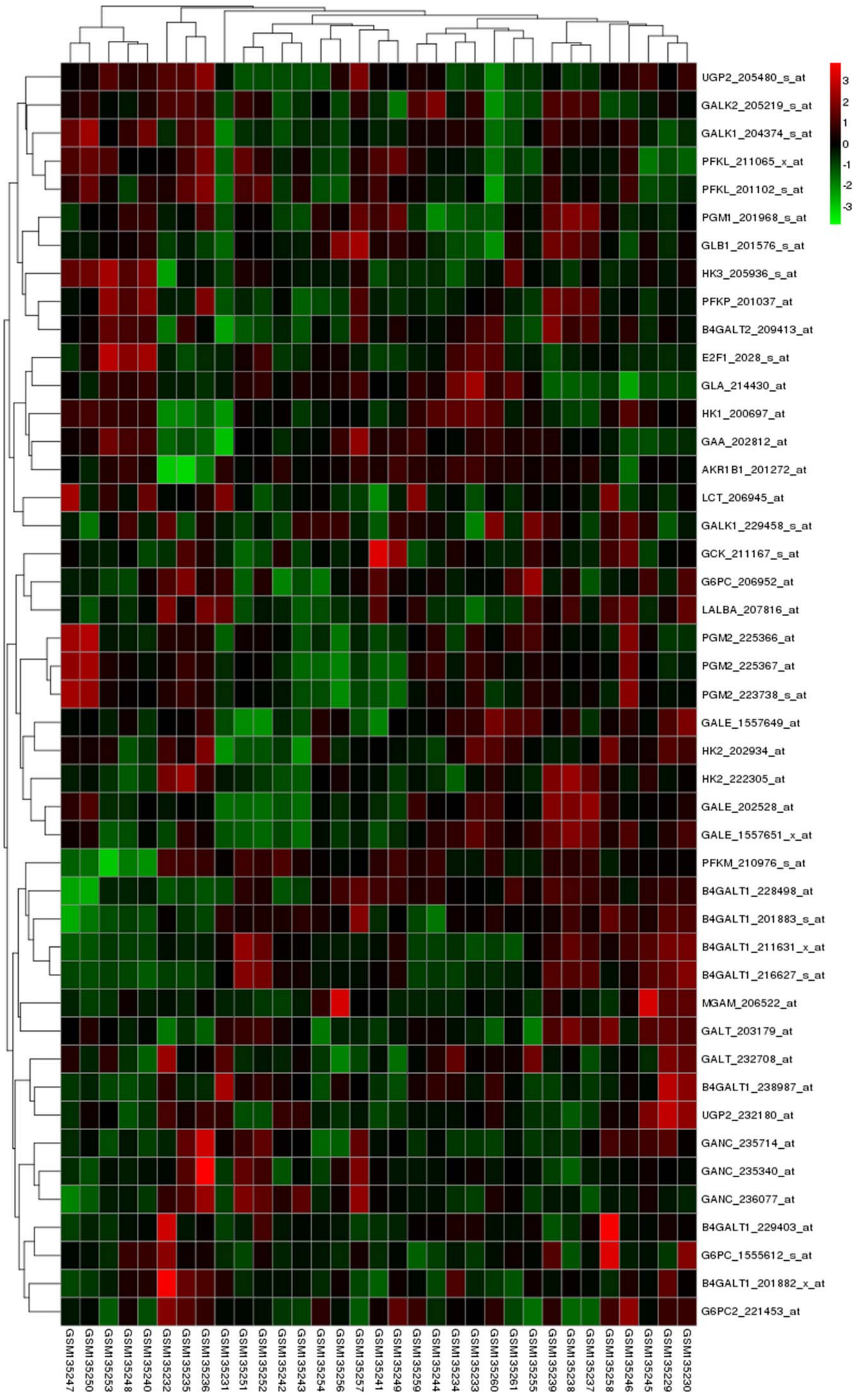

Figure 1. A heatmap showing the hierarchical clustering of the interested gene and geneset.
doi:10.1371/journal.pone.0099834.g001

contains clinical information, genomic characterization data and high level sequence analysis of the tumor genomes. The data was then log (base 2) transformed and median centred. To avoid computational error during calculation, the row that contained 'NA' value would be omitted.

The experiments were processed via various platforms (Table 1). To make the expression data searchable regardless of the platforms, the probes were remapped to official gene symbols. However, instead of gene symbol assignment information, some GPL files provided only NCBI GenBank [11] or NCBI Refseq [12] Accession Numbers mapping to probes. To solve this problem, the 'gene2refseq' and 'gene2accesion' files were retrieved from the NCBI ftp server via ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/. A Perl script was used to map gene symbols to these GenBank or RefSeq Accession Numbers, and eventually to the microarray probes. The gene expression flat files were stored for later accessing.

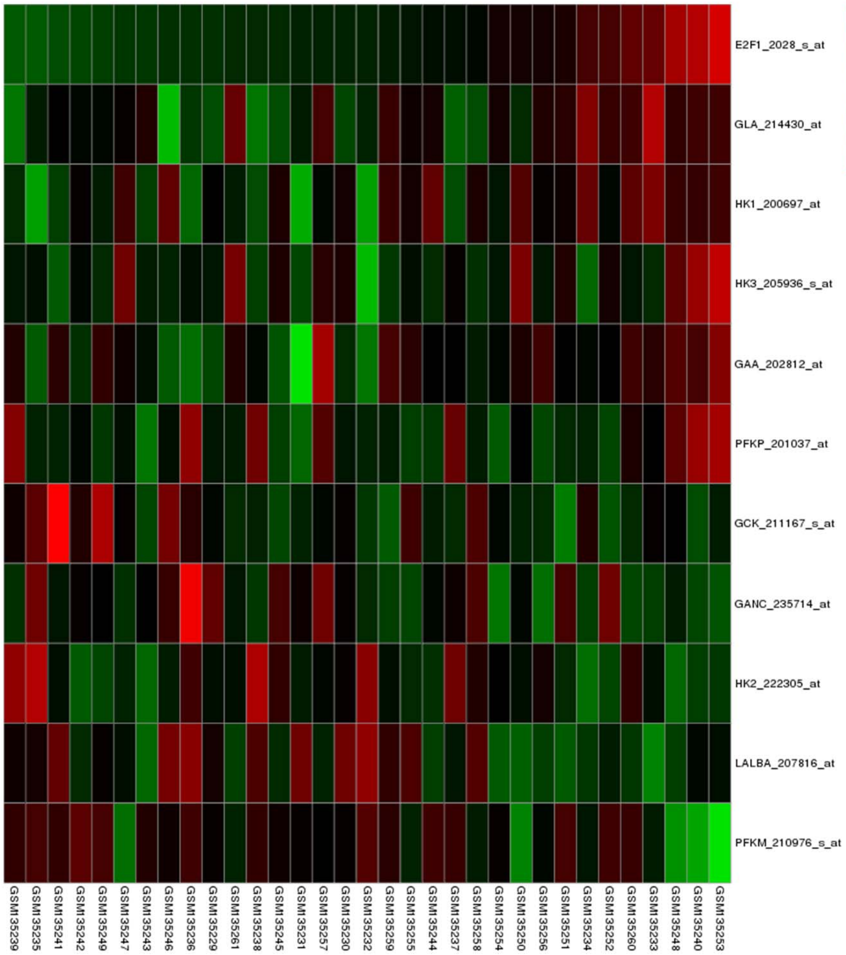

Figure 2. A heatmap showing the genes that are positively or negatively correlated with the interested gene. The genes that have significant pearson correlation with the interested gene were selected to plot a heatmap. The samplers are in the column, and ordered by the expression of the interested gene.
doi:10.1371/journal.pone.0099834.g002

Predefined Gene Set

One important feature of this database is that it enables users to search similar gene candidates with genes they are studying based on the expression patterns. Relying on this method, researchers may find mechanisms among these genes, which may become a promising approach to discovering novel gene function. The gene sets predefined in the databases were retrieved from various sources and divided into two main categories: Gene Ontology (GO) [19] and Pathway. As shown in Table 2, the GO set consists of Biological process, Molecular functions and Cellular Component. While the Pathway set consists of KEGG [20], BIOCARTA (www.biocarta.com) and REACTOME [21]. Human species of the gene sets were used in this work.

Gene Expression Cluster Analysis

The unsupervised hierarchical clustering algorithm was introduced to find the similar genes based on expression patterns. This attempt was processed using a combination of distance metrics and linkages. In this study, the distance from gene x to gene y defined as 1-r_{xy}, where r_{xy} represents the Pearson Correlation of gene x and y:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\overline{(y_i-\overline{y})}}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

Database Implementation

The dbCerEx database is a web-based utility combining a MySQL (http://www.mysql.com/) database management system [MySQL 5.5.32 (Community Server) with InnoDB engine]. The front-end web interface is enhanced by a java script framework, Bootstrap 2.3.1 (http://getbootstrap.com/). The PHP [version 5.3.10] (http://www.php.net/) applications receive the query from the user, are connected to the database to gather data, call external Perl and R scripts to process statistical analyze and generate HTML pages displaying results.

Utility and Discussion

The dbCerEx database is provided by a web-based interface. Users can start the search by entering one interested gene in the top input box, and then click on 'Search' button. A gene list will be shown in a new page for all the genes related to input gene keyword. Users can select a gene from the list according to the description to do expression analysis.

By clicking a gene, a general summary including full name, aliases and external links such as HNGC, Entrez Gene, Ensembl. MIM and Genecard for this gene will be shown. In the same page, users are allowed to set the parameters of expression analysis in cervical cancer. Users can enter an interested gene set by hand or from the gene set list such as KEGG, BIOCARTA, REACTOME and Gene Ontology. Users can select dataset from the precompiled cervical cancer expression datasets from microarray and RNASeq, or just provide a GEO accession number. By clicking the Submit Query button, the samples for the selected dataset will be listed. Users can select all or some interested samples to do expression analysis.

A heatmap displaying the hierarchical clustering of genes and samples will be shown (Figure 1). In addition, a heatmap that includes the significantly positively or negatively correlated genes with the interested gene will be also offered (Figure 2). The pearson correlation and p value will be shown as a table at the right side of the heatmap.

Conclusion

We present dbCerEx, a database containing cervical cancer gene expression profiles. In addition, it provides a novel utility for gene expression similarity search within certain interested gene sets. It is believed that dbCerEx is a powerful platform for bioinformatics discovery that brings cervical cancer microarray data and RNA-Seq data, and analysis of the cervical cancer research community with easy reach.

Availability and Requirements

The dbCerEx database website is available free of charge as a web application at: http://128.135.207.10/dbCerEx/.

Author Contributions

Conceived and designed the experiments: YH. Performed the experiments: LZ WZ ZJ YW DZ QT. Analyzed the data: LZ ML JF. Contributed reagents/materials/analysis tools: YH. Wrote the paper: LZ YH.

References

1. Walboomers J (1999) Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. The Journal of Pathology 19: 12–19.
2. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, et al. (2006) GenePattern 2.0. Nat Genet 38: 500–501.
3. Zhai Y, Kuick R, Nan B, Ota I, Weiss SJ, et al. (2007) Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion. Cancer Res 67: 10163–10172.
4. Rajkumar T, Vijayalakshmi N, Sabitha K, Shirley S, Selvaluxmy G, et al. (2009) A 7 gene expression score predicts for radiation response in cancer cervix. BMC Cancer 9: 365.
5. Imadome K, Iwakawa M, Nakawatari M, Fujita H, Kato S, et al. (2010) Subtypes of cervical adenosquamous carcinomas classified by EpCAM expression related to radiosensitivity. Cancer biology & therapy 10: 1019–1026.
6. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2013) NCBI GEO: archive for functional genomics data sets-update. Nucleic acids research 41: D991–995.
7. Team R (2008) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
8. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
9. Davis S, Meltzer P (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics.
10. The Cancer Genome Atlas (TCGA) website. Available: https://tcga-data.nci.nih.gov/tcga. Accessed: 22 May 2014.
11. Bachtiary B, Boutros PC, Pintilie M, Shi W, Bastianutto C, et al. (2006) Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity. Clin Cancer Res 12: 5632–5640.
12. Iwakawa M, Ohno T (2007) The radiation-induced cell-death signaling pathway is activated by concurrent use of cisplatin in sequential biopsy specimens from patients with cervical cancer. Cancer Biology & Therapy: 905–911.
13. Pyeon D, Newton Ma, Lambert PF, den Boon JA, Sengupta S, et al. (2007) Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. Cancer research 67: 4605–4619.
14. Kloth J, Gorter A, Fleuren G (2008) Elevated expression of SerpinA1 and SerpinA3 in HLA-positive cervical carcinoma. The Journal of Pathology: 222–230.
15. Scotto L, Narayan G (2008) Identification of Copy Number Gain and Overexpressed Genes on Chromosome Arm 20q by an Integrative Genomic Approach in Cervical Cancer: Potential Role in Progression. Genes, Chromosomes and Cancer 765: 755–765.
16. Mine KL, Shulzhenko N, Yambartsev A, Rochman M, Sanson GFO, et al. (2013) Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer. Nature communications 4: 1806.
17. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, et al. (2013) GenBank. Nucleic acids research 41: D36–42.
18. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic acids research 40: D130–135.
19. Ashburner M, Ball C, Blake J, Botstein D (2000) Gene Ontology: tool for the unification of biology. Nature genetics 25: 25–29.
20. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research 27: 29–34.
21. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic acids research 39: D691–697.